Using diversity measures for generating error-correcting output codes in classifier ensembles

Pattern Recognition Letters 26 (2005) 83–90

Ludmila I. Kuncheva

February 3, 2012

Using diversity measures for generating error-correcting output codes in classifier ensembles

Ludmila I. Kuncheva

Pattern Recognition Letters 26 (2005) 83–90

- Models
  - Error-correcting output codes
  - ECOC generation methods
- Proposal
  - Diversity measures on classifier ensembles
- Demonstration
  - Hamming distance is insuffecient for ECOC classifier ensembles.
- Methods
  - Evolutionary algorithm to construct ECOC.

- Experiments
  - Classification accuracies of 2 ensembles, using Hamming distance and Diversity measures.
- Conclusions
  - More diverse classifiers make a better ensemble than less diverse classifiers.

Outline

Introduction I

- Error-correcting output codes (ECOC) using idea: to avoid solving the multiclass problem directly and to break it into dichotomies instead.
- Example:
  - $\Omega = \omega_1, ..., \omega_{10}$ is the set of class labels.
  - We can break Ω into $\Omega = \Omega^{(1)}, \Omega^{(0)}$ where $\Omega^{(1)} = \omega_1, ..., \omega_5$ and $\Omega^{(0)} = \omega_6, ..., \omega_{10}$, called a dichotomy.
  - Discriminating between $\Omega^{(1)}$ and $\Omega^{(0)}$ will be the task of one of the classifiers in the ensemble. Each classifier is assigned a different dichotomy.
- Pressumption: diverse classifiers are obtained from diverse dichotomies.
- We propose to use diversity measures originally devised for classifiers outputs.

Outline

Error-correcting output codes (ECOC) I

- Let $\Omega = \omega_1, ..., \omega_c$ be a set of class labels .
- Suppose that each classifier codes the respective compound class $\Omega^{(1)}$ as 1 and compound class $\Omega^{(0)}$ as 0.
- Then every class $\omega_j, j = 1, ..., c$, will have a binary "profile" or a codeword.

The code matrix I

- Each dichotomy is a binary vector of length c with 1's for the classes in $\Omega^{(1)}$ and 0's for the classes in $\Omega^{(0)}$.
- Hamming distance between $[0,1,1,0,1]^T$ and $[1,0,0,1,0]^T$ is the maximum but they are identical.
- 2^c splits –> 2^{c-1}-1 splits ($\{0,\Omega\}$ is not used).

The code matrix

- Let L be the chosen number of classifiers in the ensemble.
- Class assignements: binary *code matrix* C of size c x L.
- The (i,j)th entry of C, denoted C(i,j) is 1 if class ω_j is in $\Omega_j^{(1)}$or 0, if class ω_j is in $\Omega_j^{(0)}$.
- Each row of the code matrix is a codeword and each column is a classifier assignement.

The code matrix

- Let $[s_1, ..., s_L], s \in \{0, 1\}$ be the binary output of the L classifiers in the ensemble for a given input x.
- The Hamming distance between the classifier outputs and the codewords for the classes is calculated as $\sum_{i=1}^{L} |s_i - C(j, i)|$.
- In the standard set-up the input is labeled in the class with the smallest distance (decoding phase).

The code matrix

- The code matrix should be built according to two main criteria:
  - ***Row separation:*** the codewords should be as far apart from one another as possible.
  - ***Column separation*:** dichotomies given as the assignments to the ensemble members should be as different from each other as possible too.

The code matrix

- ***Row separation:*** A measure of the quality of an error-correcting code is the minimum Hamming distance, H_c, between any pair of codewords.
- ***Column separation***: The distance between the columns must be maximized keeping in mind that the complement of a column gives the same split of the set of classes.
- Maximize:

$$H_L = \min_{i,j,i\neq j} \min \left\{ \sum_{k=1}^{c} |C(k,i) - C(k,j)|, \sum_{k=1}^{c} |1 - C(k,i) - C(k,j)| \right\}, \quad i,j \in \{1,2,\ldots,L\}. \qquad (1)$$

ECOC generation methods I

- *One-per-class*:
  - It is used as the target output for training neural network classifiers for multiple classes.
  - The target output for class ω_j is a codeword with c elements, containing 1 at position j and 0's elsewhere.
  - The code matrix is the identity matrix of size c and we only build $L = c$ classifiers.
- *All pairs:*
  - every pair of classes is taken as $\Omega^{(1)}$ and the remaining c-2 classes form $\Omega^{(0)}$.
  - There are $L = c(c-1)/2$ classifiers.
  - The minimum Hamming distance across the whole code is $2(c-2)$. The power of the all pairs code is $\left\lfloor \frac{2(c-2)-1}{2} \right\rfloor = c - 3$.

ECOC generation methods I

- *Exhaustive codes:*
  - Generating all possible $2^{(c-1)}$ different classifier assignements (for $3 \leqslant c \leqslant 7$).

1. Row 1 is all ones.
2. Row 2 consists of $2^{(c-2)}$ zeros followed by $2^{(c-1)} - 1$ ones.
3. Row 3 consists of $2^{(c-3)}$ zeros, followed by $2^{(c-3)}$ ones, followed by $2^{(c-3}$ zeros, followed by $2^{(c-3)} - 1$ ones.
4. In row i, there are alternating $2^{(c-i)}$ zeros and ones.
5. The last row is 0, 1, 0, 1, 0, 1, . . ., 0.

- *Random Generation.*

ECOC generation methods I

- Exhaustive code for $c = 4$

Exhaustive ECOC for $c = 4$ classes ($L = 7$ classifiers)

	D_1	D_2	D_3	D_4	D_5	D_6	D_7
ω_1	1	1	1	1	1	1	1
ω_2	0	0	0	0	1	1	1
ω_3	0	0	1	1	0	0	1
ω_4	0	1	0	1	0	1	0

Outline

Why is minimum Hamming distance insufficient for ECOC classifier ensembles? I

- High minimum distance between any pair of codewords implies a reduced bound on the generalization error.
- We may wish to design a code which is allowed to fail occasionally in recovering the true class label for a small number of objects but which on average will perform better than a code with a larger minimum Hamming distance.

Why is minimum Hamming distance insufficient for ECOC classifier ensembles?

Codematrix 1

	D_1	D_2	D_3	D_4	D_5
ω_1	1	0	0	0	0
ω_2	0	1	0	0	0
ω_3	0	0	1	0	0
ω_4	0	0	0	1	0
ω_5	0	0	0	0	1

Codematrix 2

	D_1	D_2	D_3	D_4	D_5
ω_1	1	0	0	0	0
ω_2	1	1	1	1	0
ω_3	1	0	1	1	1
ω_4	0	0	0	0	0
ω_5	0	1	0	1	1

H_c (min $H_c = 2$)

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	2	2	2	2
ω_2	2	0	2	2	2
ω_3	2	2	0	2	2
ω_4	2	2	2	0	2
ω_5	2	2	2	$\boxed{2}$	0

mean $H_c = 2$

H_c (min $H_c = 1$)

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	3	3	1	4
ω_2	3	0	2	4	3
ω_3	3	2	0	4	3
ω_4	$\boxed{1}$	4	4	0	3
ω_5	4	3	3	3	0

mean $H_c = 3$

95% CI for the classification accuracy

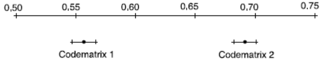

Fig. 1. An example of two ECOC ensembles. Maximizing the minimum Hamming distance will give preference to ensemble 1 which is less accurate *on average*.

Why is minimum Hamming distance insufficient for ECOC classifier ensembles? I

- According to the maximum min H_c criterion, we will prefer ensemble 1 to ensemble 2.
- A simulation was run to estimate classification accuracies of the two ensembles under the following assumptions:
  - Each of the 5 classes comes with the same probability of $1/5$.
  - Each classifier makes a mistake with probability $p = 0.2$. (A mistake here means that the 0's and the 1's in the column for the respective classifier are swapped.)

Why is minimum Hamming distance insufficient for ECOC classifier ensembles? I

- **PROCEDURE** (for 10000 objects simulated)
  1. Pick a class label with probability $1/5$. Call it ''the true label'', and denote it by i, $i \in 1, 2, 3, 4, 5$.
  2. Copy the code matrix in another matrix, C.
     1. For each classifier, decide with probability $p = 0.2$ whether it will make an error for this object.
     2. If yes, swap the 0's and the 1's in the corresponding column of C.
  3. If there were no misclassifications, the codeword for this object would be row i of the original code matrix. With the misclassifications made by the classifiers, the codeword now is the ith row of C, denotedC_i . We calculate the Hamming distances between C_i and each row of the original code matrix.

Why is minimum Hamming distance insufficient for ECOC classifier ensembles? II

4. The class label assigned by the ensemble is determined by the minimum of the five distances. In case of a tie, the assigned label is decided with equal probability between the tied labels. If the assigned label matches the true label, *i*, we increment the count for the correct classification.

- Ensemble 2 outperforms ensemble 1 by a large margin, showing that the minimum Hamming distance may not be the best criterion.

Outline

- *Dissageement measure of diversity:* between two codewords C_i and C_j is equivalent to the Hamming distance

$$\boxed{D_{i,j} = \frac{N^{01}+N^{10}}{N^{00}+N^{11}+N^{01}+N^{10}} = \frac{N^{01}+N^{10}}{L},}$$

N^{mn}:number of bits for which $C_i = m$ and $C_j = n$, $m, n \in \{0, 1\}$
L: length of the codeword

Using diversity measures for ECOC I

- If we measure column separation, the inverse of a binary vector present the same dichotomy.
- The diversity between D_i and D_j is:

$$D_{i,j} = \min\left\{\frac{N^{01}+N^{10}}{c}, \frac{N^{00}+N^{11}}{c}\right\}$$

- Total diversity between codewords:

$$D_c = \frac{2}{c(c-1)} \sum_{i<j} D_{i,j}, \quad i,\ j = 1,\dots,c.$$

Using diversity measures for ECOC I

- Total diversity between dichotomies:

$$D_L = \frac{2}{L(L-1)} \sum_{i<j} M_{i,j}, \quad i,\ j = 1, \ldots, L.$$

Using diversity measures for ECOC I

H and D for ECOC generated by the one-per-class and all-pairs methods, and for the two code matrices from Fig. 1

	Row separation (codewords)	Column separation (dichotomies)
One-per-class (=Codematrix 1)	$H_c = 2$ $D_c = \frac{2}{c}(= 0.4)$	$H_L = 2$ $D_L = \frac{2}{c}(= 0.4)$
All-pairs	$H_c = 2(c-2)$ $D_c = \frac{4(c-2)}{c(c-1)}$	$H_L = \min\{2, c-4\},\ c \geqslant 4$ $D_L = \frac{c^3-5c^2+22c-32-\lvert c-8\rvert(c^2-5c+6)}{2c(c^2-c-2)}$
Codematrix 2	$H_c = 1$ $D_c = 0.6$	$H_L = 1$ $D_L = 0.32$

- We have to combine the row and column separation measures to formulate one criterion function:
  - $D = \frac{1}{2}(D_C + D_L)$ and $H = H_C + H_L$
- We will choose **ensemble 2** because the sum is larger.

Outline

1. Introduction
2. Error-correcting output codes (ECOC)
   - The code matrix
   - ECOC generation methods
3. Why is minimum Hamming distance insufficient for ECOC classifier ensembles?
4. Using diversity measures for ECOC
5. **Generating ECOC by an evolutionary algorithm (EA)**
6. Conclusions

Generating ECOC by an evolutionary algorithm (EA)

- We use an Evolutionary algorithm to generate ECOC instead of random search.
- The chromosome is the code matrix, concatenating all rows (*Lxc*, classifiers x classes)
- **Procedure**
  - Generate Population: m chromosomes.
  - Duplicate into a offspring set.
  - Mutate each set with a specified probability P_{mut}.
  - Evaluate each chromosome
    - Breaking it, rearranging back the code matrix and calculating the chosen measure M (H or D).
  - The population and the offspring sets are then pooled and the best m of the chromosomes survive to be the next population.
  - Run these steps a number of generations.

Generating ECOC by an evolutionary algorithm (EA)

- Calculating measure: $c = 50$, $L = 15$. Parameters $m = 10$, $P_{mut} = 0.15$, num. generations =100.

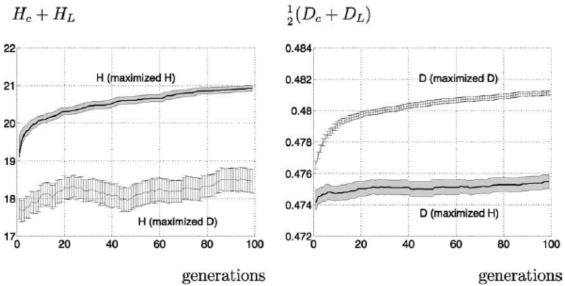

Fig. 2. *H* and *D* as functions of the number of generations in the EA (average from 100 runs; 95% confidence intervals displayed).

Outline

1. Introduction
2. Error-correcting output codes (ECOC)
   - The code matrix
   - ECOC generation methods
3. Why is minimum Hamming distance insufficient for ECOC classifier ensembles?
4. Using diversity measures for ECOC
5. Generating ECOC by an evolutionary algorithm (EA)
6. Conclusions

Conclusions

- Maximizing the minimum H is not necessarily optimal with respect to the overall correctness of the ECOC.
- An **evolutionary algorithm** was implemented to design **ECOC**s using the measures as the fitness function.
- In general **more diverse classifiers** make a better ensemble than less diverse classifiers but the relationship is not straightforward.
- Having **diverse dichotomies** does not automatically mean that the classifiers built to solve these dichotomies will be diverse.
- The goal of this study is to devise a concrete structure (**ECOC**) which can then be used in training and testing classifier ensembles.